Multi-omics : are we looking in the wrong place ?

Graham Byrnes

Biostatistics Group

International Agency for Research on Cancer
Lyon, France

Some ad-hoc thoughts on analysis

- The need to reduce dimension
- PCA/PCR, PLS, LASSO, Treelets, Signatures ?
- What do we mean by « sparse » ?
- Basis selection : linear transform, clustering, kernel transform, indexing.

The need to reduce dimension

- Whole genome methylation via chip : 450k sites : comparable to GWAS (but in GWAS, exclude rare variants)
- Via conversion and NGS : 27 million sites : comparable to whole exome sequancing (but regard only variable sites)
- GWAS still typically analysed site by site

Whole exome ?

- Too many rare variants… which are also too rare !
- Various aggregating schemes
- None really work except for highly tuned gene-specific techniques (via MAP, GVGD, SIFT)
- Big success is Mutation Signatures.

Mutation signatures

- Divide all mutations into 96 types
- Count up how many of each across the entire genome or exome
- Further reduce the dimension by Negative Matrix Factorisation (pick out un-correlated combinations)
- Characterises eg UV, Benzo-A-Pyrene, aristolochic acid, APOP-E mechanism.

Model Selection

- LASSO was going to be « the solution »
- Works very poorly for selecting correlated variables : doesn't take advantage of averaging to improve prediction (many of the problems fixable by modified versions)
- Computationally infeasible for WG
- Feature : ensures sparse models.

PCA, PCR, PLS

- Finds new basis that maximises variance per variable (or correlation)
- « No reason to suppose that the crucial information is not in the last component » - D. Cox
- Often criticised as « difficult to interpret » and « not sparse »
- If the a model with just first two PC's gave AUC=0.99, is it really not sparse ?

PCA uniqueness

- No real biological reason to suppose PCA basis will be adapted (except maybe the first PC)
- Independent processes might generate independent features, but PCA is uniquely determined by the maximum variance condition
- Infinite choice of orthogonal bases

Fourier/Wavelets/Treelets

- These are kernel transforms
- Do not map points to points, but patterns to points (and vice-versa)
- FT : time ↔ frequency
- Wavelets : nested packets of frequency bursts
- Treelets : attempt to extend wavelet ideas to un-ordered variables

Fourier example

- Imagine a toxin that caused every 15th base to be mutated
- Could be described by a single Fourier component
- Not describable by any reasonable LASSO, Ridge-regression, PCR, PLS etc model.
- Would you call it sparse ?

Fourier vs Index transform

- FT decomposes a genome-wide signal into superposition of frequencies
- Index transform transforms the original text to a list of locations of every pattern
- Or we could throw away the location and just keep the frequency of occurrence of each pattern
- This is first step of Mutation Signatures...

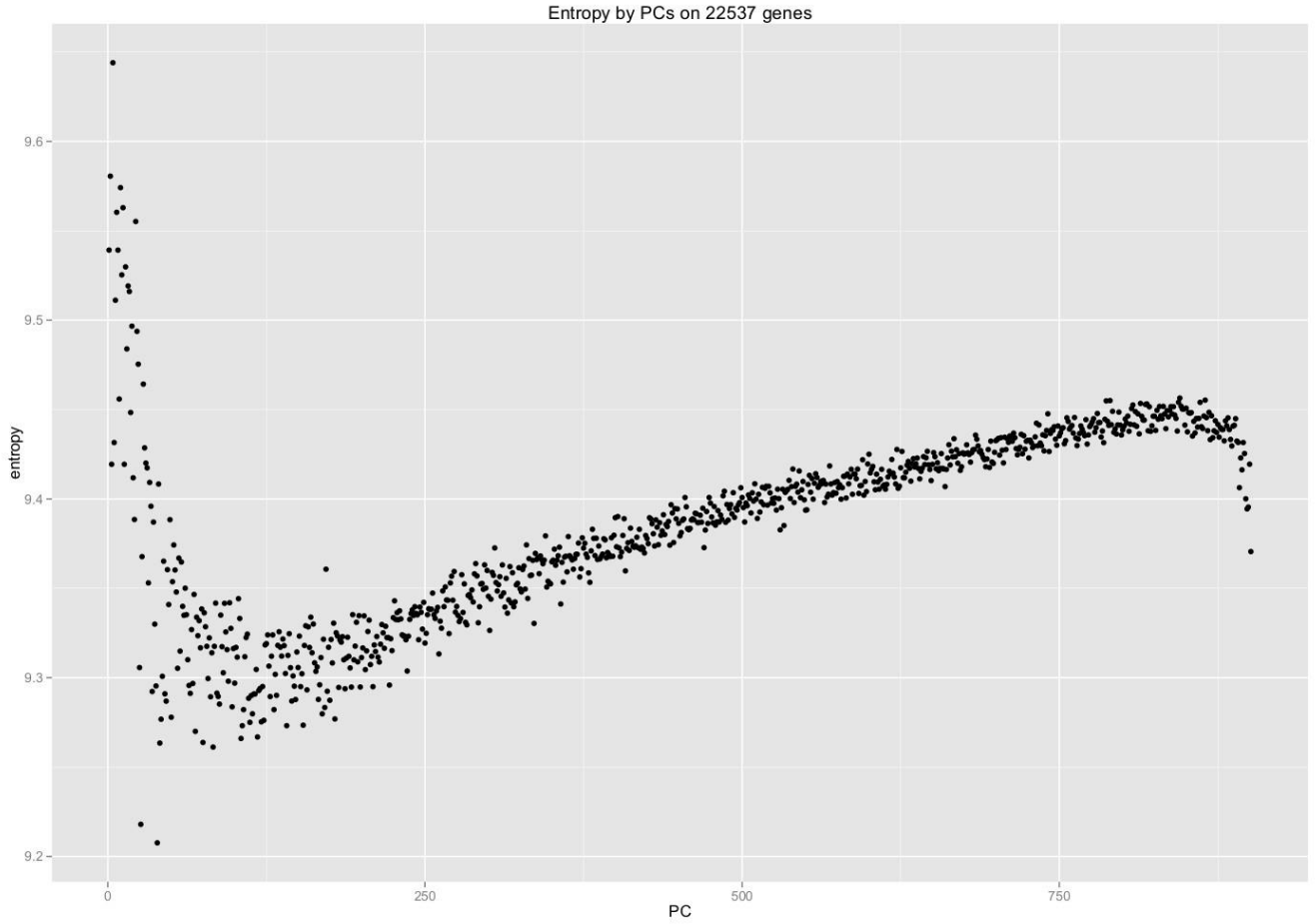
Index Transform example

- Imagine a toxin that causes A→T mutations, but only in context C(A→T)G
- Not a possible LASSO, PCR, etc model.
- Is it sparse ?
- (To make a complete transform we can add more context, ie 2, 3 or more bases 5' and 3')
- In fact IT is closest to WT.

Empirically...

- Can look to see what type of sparsity may be appropriate
- Use entropy of component loadings 22,537 genes
- Expect first PC will be close to uniform average, hence maximal entropy.
- Maximum possible Q=10.02, PC1 Q~9.65, min Q~9.25

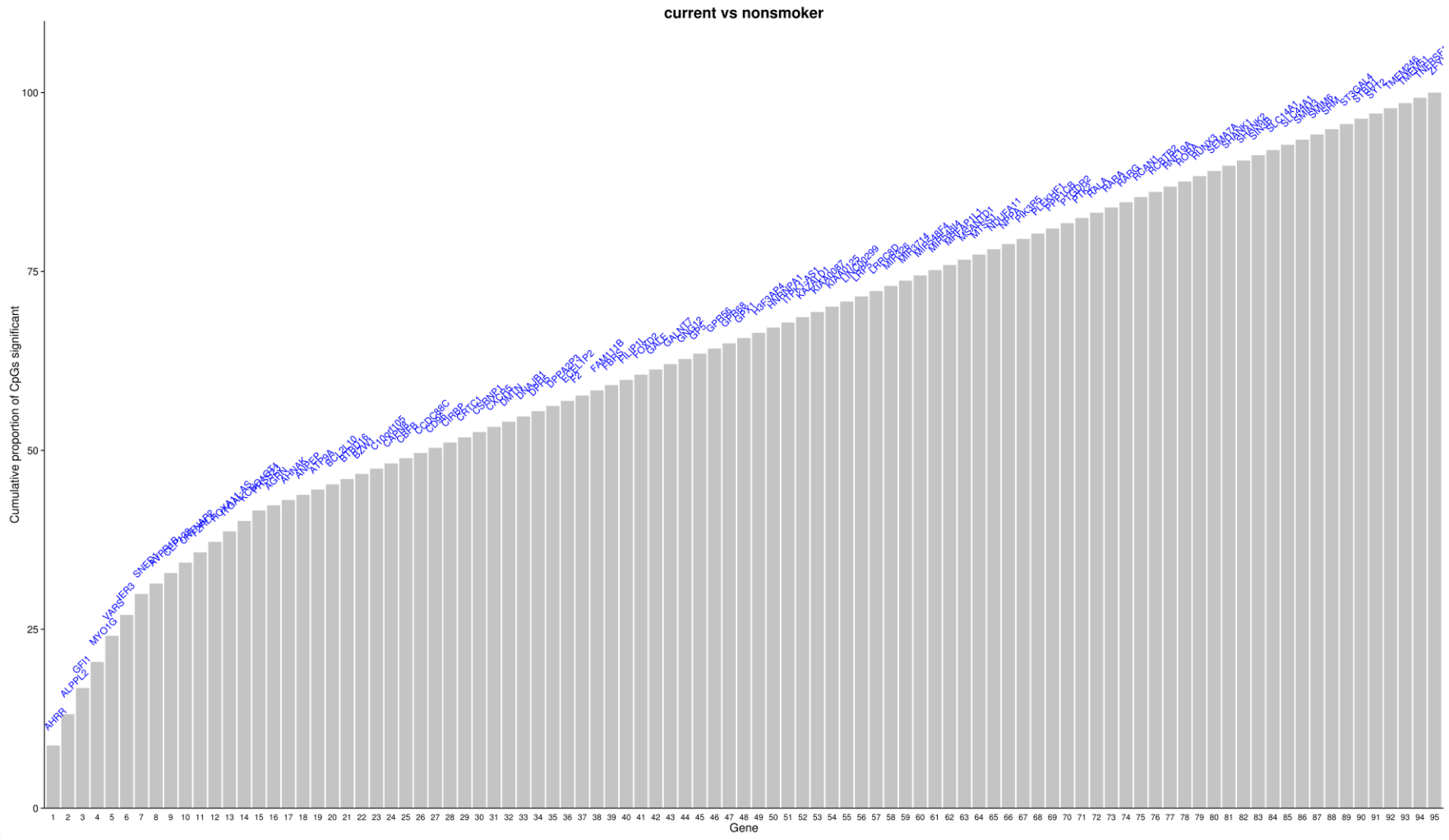
Entropy by PC



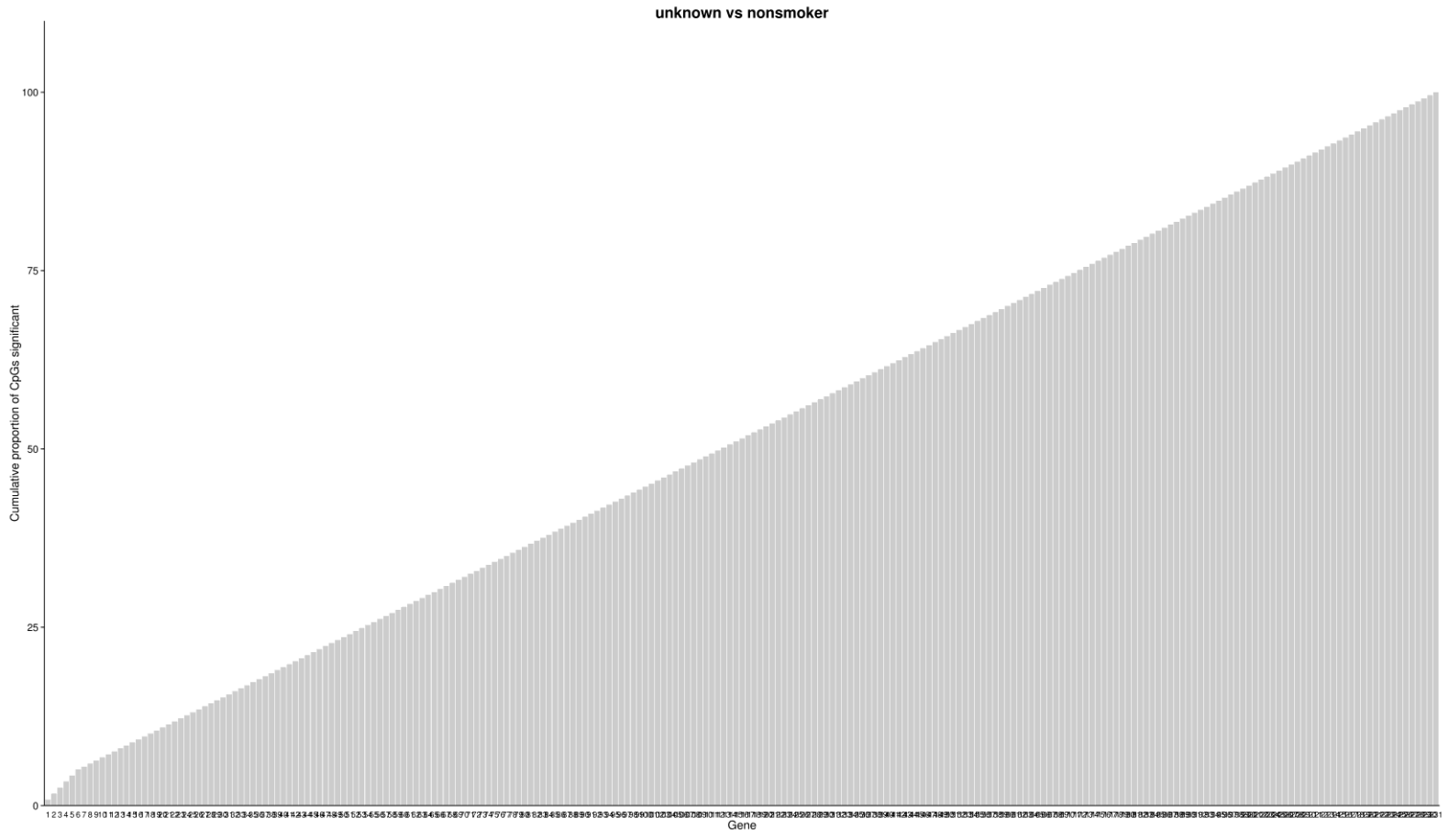
Smoking

- Use methylation data, current vs never smokers
- Take smallest p-value for each gene, 1/p re-normalise to generate a « probability of selection »
- AUC 98-99 %, 631 CPG sites FDR<0.05
- Q~1E-8

Significant CPG by gene



Unknown Smoking Status



International Agency for Research on Cancer

World Health Organization

Thanks

Liacine Bouaoun

Zdenko Herceg

Isabelle Romieu

International Agency for Research on Cancer

World Health Organization